Data Analysis Report

dataMineR

18/6/2013

CONTENTS

CHAPTER 1

INTRODUCTION

This data analysis report is generated using R-studio and knitr to knit R code and markdown into html format. We have the option to include all R code that is used to generate the plots and calculations. Default this feauture is dissabled. The data analysis step is the first step an a datamining analysis.

1.1 Dataset Basic Artifacts

Basic information from the dataset we are using.

```
i = 1

# data location full path to filename from working directory(=project dir)
# This works by default from the relative path path2file <-
# '../data/clean_base.csv'
path2file <- "../data/ano_churn_data.Rdata"

# file can be a tab delimited txt fil or a previously saveed workspace in
# .Rdata format read dataframe from tab delimited file data_set <-
# read.table('~/r-studio/NLE/data/clean_base.tab', sep= '\t', header=T,
# quote='\'') data_set <- read.delim(filename)

# read data from .Rdata save workspace
load(path2file)
# tell the script which data set to use if we load a workspace
data_set <- ano_set

# remove the original dataset
rm(ano_set)

# determine number of rows and colums in dataframe
rows <- nrow(data_set)
colums <- ncol(data_set)

# case_id = registrnr
original_case_id = "carid"
# data_set$caseID <- data_set$caseID

# check if case_id is unique
if (!(nrow(unique(data_set[original_case_id])) == nrow(data_set[original_case_id]))) {
    cat("Warning : Case_id appears not unique ! ")
}
```

```
## Error: undefined columns selected
```

We are using data from file: ../data/ano_churn_data.Rdata. The dataset has 26 variables and 92467 rows.

The case identifyer is *carid* this is unique for all cases.

1.2 Variabele types

The following variabeles are present in the dataset: age, customer_lifetime, contact_channel, product_grp, action, sales_channel, months_under_contract, P1_amount, contract_type, usage, usage2, donwpayment, extra_payment, city_code, unkown_amount, product_detail, tax_amount, P2-amount, P1_turnover, P2_turnover, index1, days_since_last_sale_p1, days_since_last_sale_p2, days_since_last_contract_adj, days_since_last_sale, target

We have 19 numeric variables and 7 categorical variables (or factors in R).

1.3 Excluded variables

From the varables provided the folowing list will be excluded in this anlysis: carid

Sometimes categoric variables are present as coded numbers. These should be treated as factors. In this dataset the following variables will be used as factors(categoric):

We have 19 numeric variables and 7 categorical variables (or factors in R).

CHAPTER 2

NUMERIC VARIABLES

Here we analyse all numeric variables. We start with an overview on basic statistics per variable. We check for missing values. We do a histogram plot to show the distribution for this variable. And we test for outliers.

2.1 Overview

In the table below we report the number of observations (n), the smallest observation (min), the first quantile (q1), the media , the mean, last quantile, the largest observation (max), and the nber of missing values (na).

```
## Attaching package: 'pander'
```

```
## The following object is masked from 'package:knitr':
##
## pandoc
```

	n obs	n missing	min	mean	median	max
age	92467	0	3.37	52.25	51.74	112.6
customer_lifetime	92467	0	1	23.93	18	92
months_under_contract	92467	0	-41	16.47	11	70
P1_amount	92467	0	0	1869	1678	77261
usage	92467	0	0	2925	2414	113879
usage2	92467	0	0	4265	3678	191185
donwpayment	92467	0	5	178.7	164	3492
extra_payment	92467	0	-14066	-46.51	-37.96	13634
city_code	92467	0	3	609.9	479	1987

unkown_amount	92467	0	1	1585	1226	12342
tax_amount	92467	0	0	0.3917	0.4114	0.475
P2-amount	92467	0	0	787	722.8	26924
P1_turnover	92467	0	0	399.2	351.2	15742
P2_turnover	92467	0	0	1186	1087	27426
index1	92467	0	0	102.8	107.3	298.9
days_since_last_sale_p1	92467	0	-1850	-221.7	28	883
days_since_last_sale_p2	92467	0	-1850	-222.8	28	883
days_since_last_contract_adj	92467	0	-137	162.9	138	1217
days_since_last_sale	92467	0	9	259.1	170	919

2.2 Variabele age

Missing: 0
Minimum value: 3.3704
Percentile 1: 24.1292
Percentile 99: 86.2369
Maximum value: 112.602

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
3.37	41.8	51.7	52.3	62.7	113

Warning : Suspect extreme values in left tail

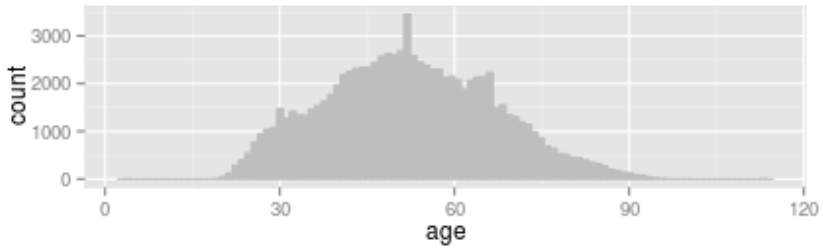


Figure 2.1: Histogram for variable age

2.3 Variabele customer_lifetime

Missing: 0
Minimum value: 1
Percentile 1: 2

Percentile 99: 67
Maximum value: 92

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	5	18	23.9	41	92

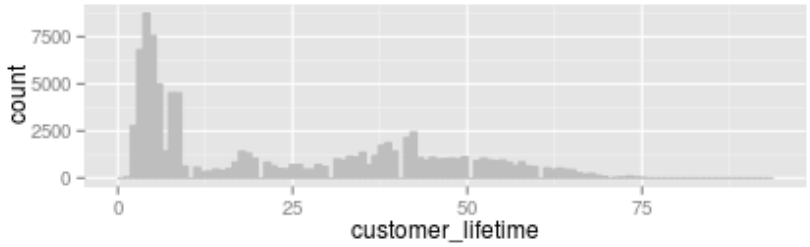


Figure 2.2: Histogram for variable customer_lifetime

2.4 Variabele months_under_contract

Missing: 0
Minimum value: -41
Percentile 1: 1
Percentile 99: 59
Maximum value: 70

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-41	9	11	16.5	16	70

Warning : Suspect extreme values in left tail

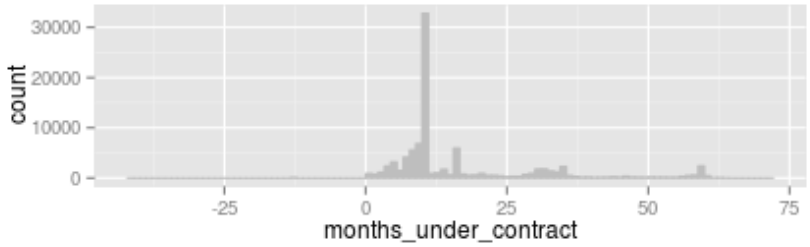


Figure 2.3: Histogram for variable months_under_contract

2.5 Variabele P1_amount

Missing: 0
Minimum value: 0
Percentile 1: 28
Percentile 99: 5914.34
Maximum value: 7.7261 × 104

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	1230	1680	1870	2250	77300

Warning : Suspect extreme values in left tailWarning : Suspect extreme values in right tail

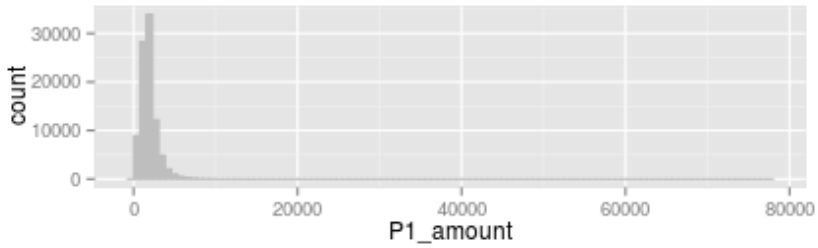


Figure 2.4: Histogram for variable P1_amount

2.6 Variabele usage

Missing: 0
Minimum value: 0
Percentile 1: 253
Percentile 99: 1.0933 × 104
Maximum value: 1.1388 × 105

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	1610	2410	2930	3580	114000

Warning : Suspect extreme values in left tailWarning : Suspect extreme values in right tail

2.7 Variabele usage2

Missing: 0
Minimum value: 0
Percentile 1: 514

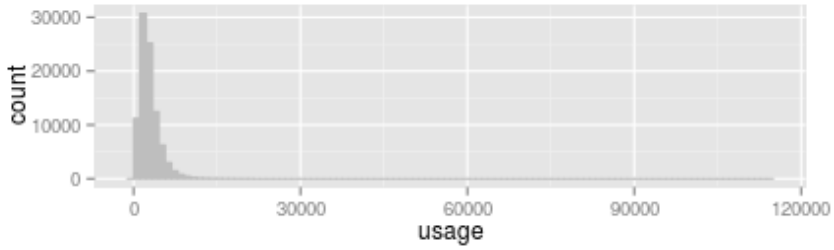


Figure 2.5: Histogram for variable usage

Percentile 99: 1.5644 × 104
Maximum value: 1.9118 × 105

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	2500	3680	4270	5130	191000

Warning : Suspect extreme values in left tailWarning : Suspect extreme values in right tail

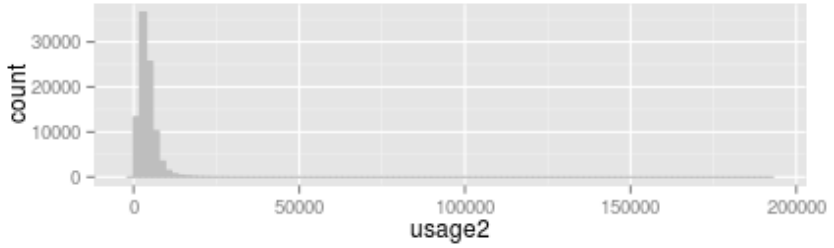


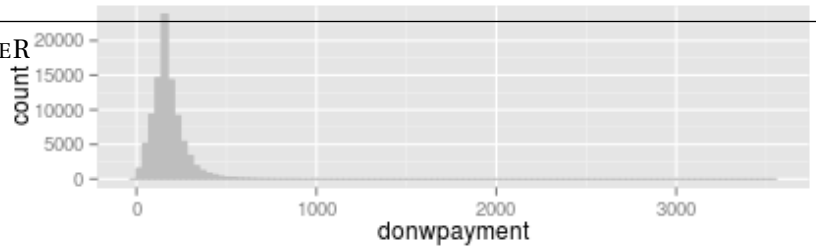
Figure 2.6: Histogram for variable usage2

2.8 Variabele donwpayment

Missing: 0
Minimum value: 5
Percentile 1: 28
Percentile 99: 530
Maximum value: 3492

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
5	123	164	179	212	3490

Warning : Suspect extreme values in left tailWarning : Suspect extreme values in right tail



2.9 Variabele extra_payment

Missing: 0
Minimum value: -1.4066 × 104
Percentile 1: -958.3522
Percentile 99: 832.5268
Maximum value: 1.3634 × 104

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-14100	-38.1	-38	-46.5	-38	13600

Warning : Suspect extreme values in left tailWarning : Suspect extreme values in right tail

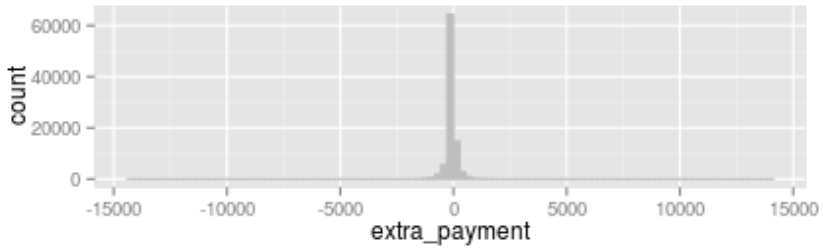


Figure 2.8: Histogram for variable extra_payment

2.10 Variabele city_code

Missing: 0
Minimum value: 3
Percentile 1: 14
Percentile 99: 1895
Maximum value: 1987

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
3	310	479	610	772	1990

Warning : Suspect extreme values in left tail

2.11 Variabele unkown_amount

Missing: 0
Minimum value: 1
Percentile 1: 25

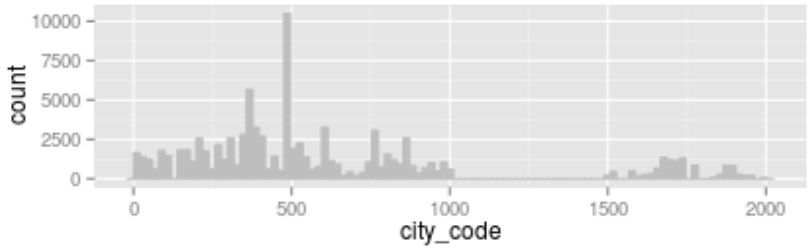


Figure 2.9: Histogram for variable city_code

Percentile 99: 8739
Maximum value: 12342

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	690	1230	1580	1930	12300

Warning : Suspect extreme values in left tail

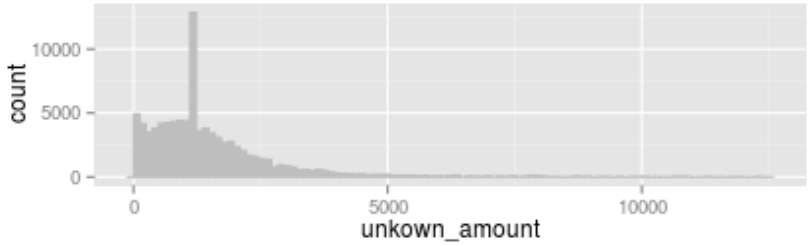
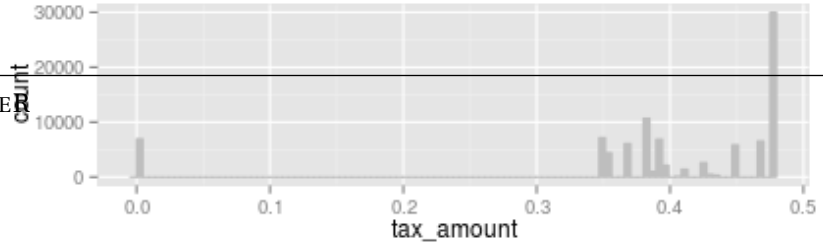


Figure 2.10: Histogram for variable unkown_amount

2.12 Variabele tax_amount

Missing: 0
Minimum value: 0
Percentile 1: 0
Percentile 99: 0.475
Maximum value: 0.475

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0.369	0.411	0.392	0.475	0.475



2.13 Variabele P2-amount

Missing: 0
Minimum value: 0
Percentile 1: 0
Percentile 99: 2599.6303
Maximum value: 2.6924 × 104

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	479	723	787	1010	26900

Warning : Suspect extreme values in right tail

```
## Error: object 'P2' not found
```

2.14 Variabele P1_turnover

Missing: 0
Minimum value: 0
Percentile 1: 0
Percentile 99: 1325.7713
Maximum value: 1.5742 × 104

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	249	351	399	482	15700

Warning : Suspect extreme values in right tail

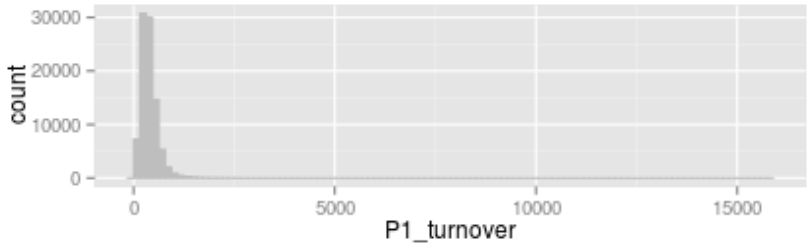


Figure 2.12: Histogram for variable P1_turnover

2.15 Variabele P2_turnover

Missing: 0
Minimum value: 0

Percentile 1: 0
Percentile 99: 3652.1753
Maximum value: 2.7426 × 104

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	766	1090	1190	1460	27400

Warning : Suspect extreme values in right tail

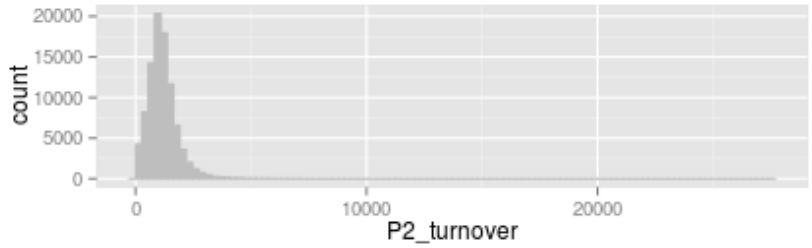


Figure 2.13: Histogram for variable P2_turnover

2.16 Variabele index1

Missing: 0
Minimum value: 0
Percentile 1: 0
Percentile 99: 176.5001
Maximum value: 298.9429

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	86.4	107	103	115	299

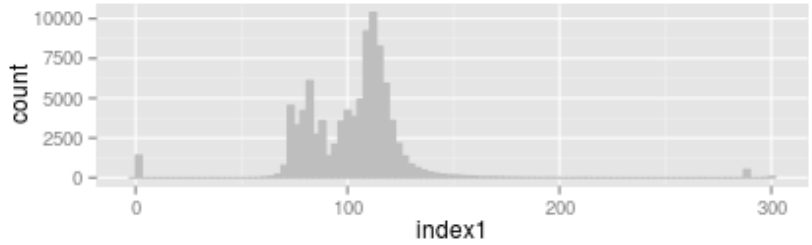


Figure 2.14: Histogram for variable index1

2.17 Variabele days_since_last_sale_p1

Missing: 0
Minimum value: -1850
Percentile 1: -1688.34
Percentile 99: 215.34
Maximum value: 883

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1850	-427	28	-222	35	883

Warning : Suspect extreme values in left tailWarning : Suspect extreme values in right tail

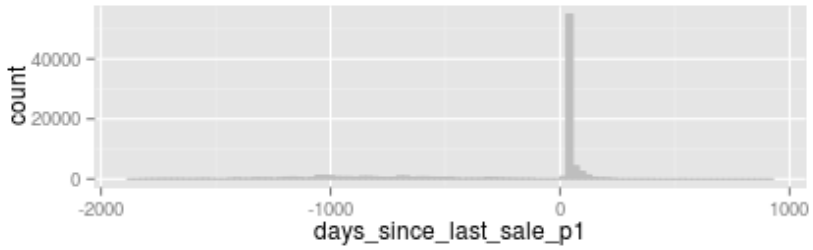


Figure 2.15: Histogram for variable days_since_last_sale_p1

2.18 Variabele days_since_last_sale_p2

Missing: 0
Minimum value: -1850
Percentile 1: -1687
Percentile 99: 225
Maximum value: 883

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1850	-450	28	-223	35	883

Warning : Suspect extreme values in left tailWarning : Suspect extreme values in right tail

2.19 Variabele days_since_last_contract_adj

Missing: 0
Minimum value: -137
Percentile 1: 18

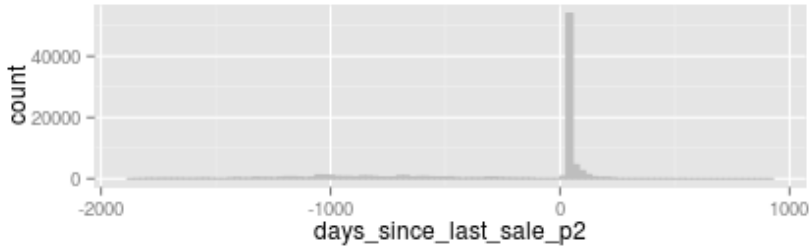


Figure 2.16: Histogram for variable days_since_last_sale_p2

Percentile 99: 389
Maximum value: 1217

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-137	77	138	163	228	1220

Warning : Suspect extreme values in left tailWarning : Suspect extreme values in right tail

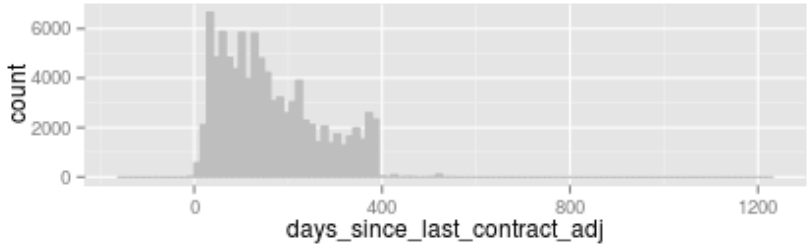
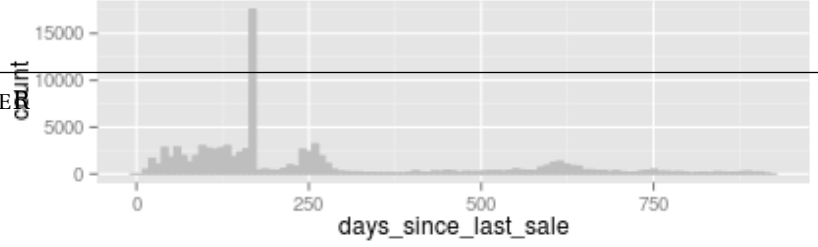


Figure 2.17: Histogram for variable days_since_last_contract_adj

2.20 Variabele days_since_last_sale

Missing: 0
Minimum value: 9
Percentile 1: 21
Percentile 99: 876
Maximum value: 919

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
9	114	170	259	281	919



CHAPTER 3

CATEGORICAL VARIABLES

Here we analyse all categorical variables. We first check the number of different levels in each category(or factor). Then we do a bar plot to show the distribution for each variable.

3.1 Overview

In the following table we will see each variable printed with it's unique levels. Beside each level a count is made and a precentage calculated. In the last colum we find a culumative count summing the total up to 100%.

We see that the number of levels can be quite big, for reporting we will omit all variables with more then 32 levels. These will not be reported in the subsections below.

	levels	missings
contact_channel	32	0
product_grp	32	0
action	32	0
sales_channel	32	0
contract_type	7	0
product_detail	32	0
target	2	0

Variables with to many levels to report are : no variabes to report .

3.2 Variabele contact_channel

The table shows the number of observations of each level.

	count
C_1	12454
C_2	1107
C_3	562
C_4	537
C_5	3462
C_6	1466
C_7	1519
C_8	10861
C_9	904
C_10	1772
C_11	2814
C_12	718
C_13	4604
C_14	1851
C_15	1453
C_16	623
C_17	1011
C_18	734
C_19	3306
C_20	532
C_21	1138
C_22	2735
C_23	3212
C_24	2437
C_25	889
C_26	950
C_27	1363
C_28	3911
C_29	626
C_30	1689
C_31	19435
C_32	1792

3.3 Variabele product_grp

The table shows the number of observations of each level.

	count
P_1	26003
P_2	12278
P_3	2449
P_4	1544
P_5	2944
P_6	2019
P_7	697
P_8	971
P_9	3004
P_10	2105
P_11	593
P_12	1395
P_13	5824
P_14	4888
P_15	610
P_16	1970
P_17	4907
P_18	13311
P_19	2141
P_20	640
P_21	1416
P_22	758
P_23	0
P_24	0
P_25	0
P_26	0
P_27	0
P_28	0
P_29	0
P_30	0
P_31	0
P_32	0

3.4 Variabele action

The table shows the number of observations of each level.

	count
A_1	14029
A_2	11933
A_3	599
A_4	486
A_5	875
A_6	576
A_7	857
A_8	5621
A_9	469
A_10	463
A_11	1707
A_12	740
A_13	1105
A_14	563
A_15	10740
A_16	1566
A_17	7226
A_18	1008
A_19	1617
A_20	1219
A_21	1318
A_22	13167
A_23	604
A_24	810
A_25	1463
A_26	5410
A_27	1059
A_28	3262
A_29	1390
A_30	585
A_31	0
A_32	0

3.5 Variabele sales_channel

The table shows the number of observations of each level.

	count
S_1	585
S_2	1109
S_3	3500
S_4	456
S_5	1275
S_6	563
S_7	11485
S_8	1034
S_9	10963
S_10	714
S_11	3478
S_12	1223
S_13	616
S_14	5715
S_15	639
S_16	2818
S_17	873
S_18	1869
S_19	821
S_20	3377
S_21	2248
S_22	984
S_23	1569
S_24	13521
S_25	1089
S_26	19943
S_27	0
S_28	0
S_29	0
S_30	0
S_31	0
S_32	0

3.6 Variabele contract_type

The table shows the number of observations of each level.

	count
T_1	597
T_2	35519
T_3	33061
T_4	6249
T_5	14556
T_6	2348
T_7	137

3.7 Variabele product_detail

The table shows the number of observations of each level.

	count
P_1	13780
P_2	2521
P_3	7856
P_4	693
P_5	2553
P_6	1932
P_7	887
P_8	1467
P_9	1244
P_10	1316
P_11	1364
P_12	5353
P_13	1198
P_14	969
P_15	631
P_16	699
P_17	2188

P_18	931
P_19	1680
P_20	2285
P_21	5195
P_22	1038
P_23	7151
P_24	1020
P_25	1072
P_26	10307
P_27	4404
P_28	795
P_29	752
P_30	854
P_31	1394
P_32	6938

3.8 Variabele target

The table shows the number of observations of each level.

	count
N	76680
Y	15787

CHAPTER 4

BEHAVIOURAL ANALYSIS

The next step is behavioural analysis. The current dataset is now saved.

Dataset saved as : ../data/data-set.Rdata